Innovative Computing Review (ICR)

Volume 1 Issue 1, Summer 2021

Homepage: https://journals.umt.edu.pk/index.php/icr

Article:

Hate Speech Detection in Social Media Surveillance: A Review of Related Literature

Journal QR

Article QR

Usman Ahmed

Author(s): Usman Ahmed[1], Rahmet Bibi[2], Obaid Ullah[2], Rahat Bano[2]

Affiliation:
[1]Foundation University Islamabad, Pakistan
[2]Superior University Lahore, Pakistan

Article History:
Received: March 14, 2021
Revised: May 19, 2021
Accepted: June 30, 2021
Available Online: June 30, 2021

Citation: U, Ahmed, B, Rahmet, U, Obid, and B, Rahat, “Hate speech detection in social media surveillance: A review of related literature”, *Innova Comput Rev*, vol. 1, no. 1, pp. 01–11, 2021.

A publication of the

School of Systems and Technology

University of Management and Technology, Lahore, Pakistan

Hate Speech Detection in Social Media Surveillance: A Review of Related Literature

Usman Ahmed[1], Rahmet Bibi[2], Obaid Ullah[2], Rahat Bano[2]

ABSTRACT: Social media surveillance is a requirement for governments and intelligence agencies around the world to detect and prevent hate crimes. The dynamic and unstructured nature of the textual content available on social media platforms makes it very complex to extract hate related speech patterns from this content. It also creates ambiguities in the data and therefore, data mining techniques become difficult to apply in this scenario. Several alternative techniques were adopted by different researchers in the past to cope with this problem and to capture and analyze such unstructured text for the purpose of hate speech detection. In this paper, we reviewed, categorized and presented a state-of-the-art of these techniques which were divided into three categories namely text mining, sentiment analysis and semantics. The challenges in the application of the existing techniques were also discussed and these can be taken up as future directions.

INDEX TERMS: data mining, hate speech, NLP, semantics, sentiment analysis, social media, surveillance, text mining

I. Introduction

Today, social media platforms have made it possible for people around the world to connect and communicate with each other, easily. However, it has also made it easy for criminals to perform criminal activities using these platforms. Governments and intelligence agencies need to detect and prevent such criminal activities using data mining and other similar techniques on the data collected from these sources. Although, there has been an extensive debate about the amount of private information that can be gathered for surveillance by the governments and agencies [1].

Social media platforms such as Facebook, Twitter and LinkedIn provide extensive opportunities for the masses to mutually connect and engage in order to share knowledge and to communicate with each other. Textual data communicated over social media platforms in the form of comments, messages, and posts remains mostly

[1] Usman Ahmed is with Foundation University Islamabad, Pakistan.Usman Ahmed is corresponding author and available at usman.ahmed@fui.edu.pk

[2] Rahmet Bibi, Obaid Ullah, and Rahat Bano are with Superior University Lahore, Pakistan.

unstructured because people are more likely to use language without proper spellings and grammar. It creates ambiguity in sentences and words and consequently, semantic and syntactic problems while interpreting this data and uncovering logical patterns from it [2].

The problems concerning the ambiguities in the data can be resolved using textual mining or simply text mining. Although uncovering patterns from the unstructured data is a challenge because it involves searching and analyzing the structured data found in the database to find the patterns of interest. Text mining, on the other hand, is a blend of techniques including natural language processing (NLP), text analysis, and information retrieval. It is used to find out logical patterns from the unstructured data [3]. Text mining is more complex than data mining because of the nature of natural language. Moreover, social media is replete with unstructured textual content in its posts and comments.

Several approaches to text mining are found in the literature. However, there still exist some major challenges related to text mining in social media including the fact that the data is much larger and dynamic in nature. Moreover, there are the privacy concerns. The use of semantic approaches for coping with such problems in social media is also a recent trend. It is the approach of giving meaning to the terms and concepts to find relationships among them that uncover patterns of interest. Ontologies are explicit and bring forth the shared conceptualization of real-world scenarios [4]. They provide machine-processable semantics, therefore, are used in different domains such as software engineering, medical science, and others to automate the different tasks involved. Ontologies are also used in text mining in social media analysis [5]. Several approaches are used to employ ontologies to infer racial, religious or political hate speech using the content shared on social media platforms in the form of posts and comments. The current paper is organized as a literature review about the previous techniques and divides them into the domains of text mining, sentiment analysis and semantics, followed by a state-of-the-art of these techniques. Lastly, there are the conclusion and future directions.

II. Literature Review

Social networking platforms such as Facebook and Twitter are used to remove barriers among communities around the world and to share and communicate knowledge and thoughts [6]. Unfortunately, these platforms have also enabled people to commit insults, cyberbullying, and various incitements through racial, religious and political hate speech that lead to social anarchy [7].

Hate speech as a term was used by Werner [8] and it is defined as an intentional act of insult using abusive or hostile messages. Several other terms are used in the literature including cyberbullying for the same purpose. The phenomenon of hate speech has negative consequences in the society and remains punishable around the world by governments using different laws [9], [10]. It is thus necessary for any government to detect and respond to this crime through surveillance. Here, we are concerned with surveillance involving textual data from various social media websites.

Several techniques are used for analyzing unstructured content on social media. Social media platforms are loaded with huge amounts of textual data. The communication of the majority of social media users takes the form of unstructured data, not using the exact words and proper sentence structure. Although data mining techniques have been used to extract huge amounts of structured data, these techniques can’t work well with the unstructured and dynamic textual data. Therefore, text mining and natural language processing techniques have become popular for analyzing unstructured textual data. Text mining is an extension of data mining but it is more complex and different because of its dealing with natural language. It creates meaningful data from unstructured data patterns [2].

Text mining or text analytics can be divided into four stages. Firstly, it involves capturing the data from social media platforms and pre-processing it using stemming, tokenization and stop-word removal techniques. The extracted and cleaned data is then represented through models to derive knowledge from it. For the representation and modeling of the data, mostly BOW (Bag of Words) technique is employed as described in [11]. It transforms the data collected from these mediums into numeric vector-based representations onto which algebraic operations can be applied. However, this technique has a shortcoming as well when it is used to analyze the relationship between various isolated pieces of information. This semantic gap can only be filled if we use the semantic knowledge base. Another text mining approach [12] extracts intellectual data from Facebook. It uses Facebook API to extract the data about several attributes of users including their age, profile, comments, and timeline posts in order to transform them into different representations using data mining techniques. The aim is to study the users themselves, their activities and to create their individual profiles. Also, this approach is concerned with the inter-user comparison of their activities and behavior prediction.

Yet another technique [13] uses Twitter tweets to identify medication abuse related posts and to monitor them.

It identifies and organizes the tweets that mention these abuse related medications and then divides them into three categories of drug or medication abuse. The authors also developed a classification technique that is automatically supervised. It distinguishes the posts that signal the presence of drug abuse and those that do not. NLP and the WordNet semantic ontology are also employed to analyze the posts. For the automated supervised classification, several available algorithms such as Support Vector Machines (SVM) and others are used. Then, all classifiers are further combined from these algorithms using stacking for making the final decision based on individual predictions.

A semantic approach by Mika [14] discusses folksonomy as a dynamic type of ontology related to a particular community-based domain (social network). Folksonomy allows the individual user to express shared concepts using its own choice of keywords. The author argues that the very basic and simple concept of ontology restricts folksonomy from being used across the domain while the problems may arise with the temporal extension of knowledge and the evolution of the social community with its changing members or through change in their commitments. This can invalidate the knowledge contained in that particular ontology. The author presented a tripartite model consisting of the actor, concept, and instance for the purpose of folksonomy. This idea was inspired by the social tagging mechanism where the user uses different tags to communicate its concerns. In this way, the particular social media context is focused on the construction of such ontology. For instance, Flickr is used to share photos, CiteULike provides scientific papers' tagging and so on. Although their scope is limited to a single website, the idea of tagging can be reused in other similar applications.

Mynard et al. [15] presented a real-time semantic framework employing linked open data as a knowledge base to use in their semantic searches. They used the GATE-based opensource framework for searching and aggregation to analyze the textual content in Twitter tweets, highlighting them in a generic or domain independent environment using specific keywords. They employed this framework in the domain of political science for the study of change in the political environment and it led them to successfully predict UK general elections. However, its main focus is not a specific domain; rather, it presents a more generic form of analysis. Also, it's limited to Twitter only.

In [16], the authors used neural language models to detect hate speech from comments. They moved away from the BOW approach as it may not capture all hate speech if the offender changes the offensive words and yet be clear in its meaning. Therefore, they used the low dimensionality approach of CBOW (Continuous Bag of Words) for the neural language model to become more effective and efficient in hate speech detection. In this approach, words are represented in a common vector space and the analysts try to predict a central word or comment.

There is another technique known as sentiment analysis that attempts to detect the sentiment or emotions of the users from the text in order to determine their attitude or opinion regarding a certain topic. This may be thought of as similar to text analysis. However, it is mostly used to determine if some expression or sentence is positive, negative or neutral. It is also used to rate a product or service on the basis of these reviews. One such approach used for analyzing social media content employed unsupervised lexicon based classification [17]. The unsupervised classifier does not require any training and therefore produces a positive and negative response rating for a given expression because both positive and negative emotions may be present in a text. So, this approach makes a ternary prediction based on the input of the estimates of both positive and negative emotions and whichever's value overwhelms the other, becomes the answer in the final prediction. Authors demonstrated that their approach produces better results as compared to the available supervised or machine learning solutions.

Another recent empirical study made the use of semantic feature representations to better understand the context of user expression in order to identify the hate speech intent of the user [18]. This approach used the external knowledge base of semantics in feature representations unlike the previous approaches. Semantic features that were employed included Hatebase features and FrameNet features. Hatebase is a multilingual hate speech knowledge base available online. The approach used several feature vectors (H_x) such as hateful meaning, non-hateful meaning, offensiveness, and unambiguousness to average such vectors in order to generate knowledge base features of a given social media post. FrameNet is another linguistic knowledge base that provides meanings under different semantic frame categories. Each post is processed through such a frame semantic parser and a vocabulary is formulated. This vocabulary is then used to create the vector for each post representing the count of each frame in a given post. Although the FrameNet approach has improved hate speech detection yet it is presently limited to English language only and it is also partially affected by the words with multiple connotations.

The state-of-the-art table of various data mining techniques is presented below in Table 1 and Table 2. We

TABLE 1

CATEGORIZATION OF TECHNIQUES USED FOR TEXT ANALYSIS IN SOCIAL MEDIA

Author	Year	Technique	Category	Main Idea	Social Network
Rahman	2012	Systematic Mining Model	Text Mining	User Attributes	Facebook
Djuric et al.	2015	Continuous Bag of Words (CBOW)	Text Mining	Low Dimensional Text Embeddings	Generic
Sarker et al.	2016	Automated Supervised Classification	Text Mining	Combining Classifiers	Twitter
Paltoglou et al.	2012	Unsupervised Lexicon Classification	Sentiment Analysis	Linguistics Functions Classifier	Twitter, MySpace, Digg
Mika	2007	Folksonomy, tripartite semantic model	Semantics	Social Tagging	Generic
Mynard et al.	2017	GATE semantic framework	Semantics	Keyword Search	Twitter
Senarath et al.	2020	Semantic features	Semantics	Declarative Knowledge Based Semantic Features	Twitter

divided all such techniques previously used for the analysis of the text in social media into three categories namely text mining, sentiment analysis, and semantics. Certain parameters including main idea, search, discovery, prediction, and methodology validation were chosen to draw a comparison among different techniques.

Table 1 shows the categories in which these works fall and also depicts the platform on which the current work is carried out. Techniques and the main idea show the adopted approach.

Table 2 shows the comparison among the parameters we have specified, that is, search that corresponds to the searching of words and phrases from the social media and

TABLE 2
COMPARISON OF TECHNIQUES W.R.T SELECTED PARAMETERS

Technique	Category	Search	Discovery	Prediction	Validation
Systematic Mining Model	Text Mining	To a certain extent	Yes	To a certain extent	No
Continuous Bag of Words (CBOW)	Text Mining	Yes	Yes	To a certain extent	Yes
Automated Supervised Classification	Text Mining	Yes	Yes	Yes	No
Unsupervised Lexicon Classification	Sentiment Analysis	Yes	Yes	To a certain extent	To a certain extent
Folksonomy, tripartite semantic model	Semantics	Yes	Yes	No	Yes
GATE semantic framework	Semantics	Yes	Yes	To a certain extent	Yes
Semantic features	Semantics	Yes	Yes	No	Yes

discovery that corresponds to the identification of the true intended meaning of the text as hate speech. Another parameter is prediction which refers to the future prediction of such actions using the given text or predicting other contents or intentions of the user by analyzing its contents. Lastly, the last column of the table indicates the validation status of the technique.

These studies provide valuable insights into the approaches and strategies used in the past for natural text analysis of the social media content. The works mentioned here also correspond to approaches and techniques adopted by several other researchers. It has been noted in the current study that most approaches use some keywords related to a specific domain, for instance hate speech in this scenario, as a targeted search across social media posts. This approach is mostly observed in the text mining technique. Some researchers [12] have adopted the approach of profiling the activities of individual users to detect hate speech. Others [16] tried to model the captured data through graphs and other representations, so as to apply mathematical operations to detect or predict hate speech using partial content. Another interesting but complex approach is to combine the results of multiple classifiers to give a final one as an automated yet supervised classification. The idea of detecting sentiments and emotions as in [17] can be very useful in hate speech detection. However, what if the attacker uses sarcastic language that may not contain abuse related words but still attacks some group or individual using sarcasm? This problem can only be dealt with if we analyze the relationship of words and sentences as in ontologies. Ontologies have also been used as semantic approaches for the analysis of textual content.

In the field of semantics, most approaches by different researchers use linked open data for identifying hate speech and analyzing the text from social media as in [15], [19]. Although this approach is helpful partially, such data can only provide us with the structured knowledge base that might not be successful in most of the cases where the data obtained from the posts and comments is unstructured, yet incorporates hate speech and abusive or slang language. An ontology can be built that may represent these unstructured terminologies in a structured and semantic manner as a knowledge base that can reuse structured concepts from linked open data as well. Currently, there is no generic ontology on hate speech.

III. CONCLUSION

Hate speech on social media has been the cause of several catastrophes and social anarchy in the past. The powerful impact of social media and our reliance on them has increased and so have the chances of hate speech. Textual surveillance by governments is therefore necessary to detect and prevent hate crimes. Most previous approaches in data mining can’t work well with the ambiguous and unstructured data emerging from social media platforms. Several techniques have been used to solve this problem. We have presented in this paper a review of the various types of techniques used in the effort to detect and analyze textual content on social media platforms. These techniques can be divided into text mining, sentiment analysis, and semantics. Each has its own advantages and shortcomings. We argued that analyzing responses or comments can provide insights into the presence and severity of hate-related social media posts. The use of semantic techniques has proven to be more efficient. However, there is still a need to explore the features of semantic web technologies for the detection of hate speech in social media, in particular to cope with the problems of multiple interpretations and sarcasm. The state-of-the-art presented in this paper can help the researchers to identify problems and present better solutions in the future.

REFERENCES

[1] D. Roark, "The end of privacy for the populace, the person of interest and the persecuted," *Health and Technology*, vol. 7, no. 4, pp. 501–517, 2017.

[2] R. Irfan, *et al.*, "A survey on text mining in social networks," *The Knowledge Engineering Review*, vol. 30, no, 2, pp. 157–170, 2015. https://doi.org/10.1017/S0269888914000277

[3] R. Feldman, and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[4] N. Guarino, *Formal ontology in information systems*. IOS Press, 1998.

[5] A. Hotho, R. Jäschke, K. Lerman, "Mining social semantics on the social web," *Semantic Web*, vol. 8, no. 5, pp. 623–624, 2017. http://dx.doi.org/10.3233/SW-170272

[6] L. Sorensen, "User managed trust in social networking - Comparing Facebook, MySpace and Linkedin," In: *2009 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace Electronic Systems Technology*, 2009, pp. 427–431.

[7] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 2017, pp. 86-95.

[8] W. Warner, and J. Hirschberg, "Detecting Hate Speech on the World Wide Web," In: *Proceedings of the second workshop on language in social media*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 19–26.

[9] B. Parekh, "Hate Speech," *Public Policy Research*, vol. 12, no. 4, pp. 213–223, 2006. https://doi.org/10.1111/j.1070-535.2005.00405.x

[10] R. M. Simpson, "Dignity, harm, and hate speech," *Law and Philosophy*, vol. 32, pp. 701–728, 2013. https://doi.org/10.1007/ s10982-012-9164-z

[11] X. Hu, and H. Liu, "Text analytics in social media," In: *Mining text data*, C. C. Aggarwal, and C. Zhai, eds. Springer US, Boston, MA, 2012, pp. 385–414. https://link.springer.com/chapter/10.1007/978-1-4614-3223-4_12

[12] M. M. Rahman, "Mining social data to extract intellectual knowledge," *International Journal of Intelligent Systems and Applications*, vol. 4, no. 10, 2012. https://doi.org/10.5815/ijisa.2012.10.02

[13] A. Sarker, *et al.*, "Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter," *Drug Saf*ety, vol. 39, no. 3, pp. 231–240, 2016. https://doi.org/10.1007/s40264-015-0379-4

[14] P. Mika, "Ontologies are us: A unified model of social networks and semantics," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 1, pp. 5–1, 2007. https://doi.org/10.1016/j.websem.2006. 11.002

[15] D. Maynard, I. Roberts, M. A. Greenwood, D. Rout, K. Bontcheva, "A framework for real-time semantic social media analysis," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 44, pp. 75–88, 2017. https://doi.org/10.1016/j.wasman.2017.08.009

[16] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, "Hate speech detection with comment embeddings," In: *Proceedings of the 24th International Conference on World Wide Web*, ACM, New York, NY, USA, 2015, pp. 29–30. https://doi.org/10.1145/2740908.2742760

[17] G. Paltoglou, and M. Thelwall, "Twitter, MySpace, and Digg: unsupervised sentiment analysis in social media," *ACM* Transactions on Intelligent Systems and Technology, vol. 3, no. 4, pp. 1–66, 2012. https://doi.org/10.1145/2337542.2337551

[18] Y. Senarath, and H. Purohit, "Evaluating Semantic feature representations to efficiently detect hate intent on social media," In: *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, Irvine, CA, USA, 2020, pp. 199–202. https://doi.org/10.1109/ICSC.2020.00041

[19] F. Sahito, A. Latif, W. Slany, "Weaving Twitter stream into Linked Data a proof of concept framework," In: *2011 7th International Conference on Emerging Technologies*, Islamabad, Pakistan, 2011, pp. 1–6. https://doi.org/10.1109/ICET.2011.6048497